Gene expression variability in mammalian embryonic stem cells using single cell RNA-seq data

Gene expression variability in mammalian embryonic stem cells using single cell RNA-seq data

Anna Mantsoki[1], Guillaume Devailly[1], Anagha Joshi[1§]

[1]The Roslin institute, University of Edinburgh, Easter bush campus, Midlothian, EH25 9RG.

[§]Corresponding author

Email addresses:

AM: Anna.Mantsoki@roslin.ed.ac.uk

GD: Guillaume.Devailly@roslin.ed.ac.uk

AJ: Anagha.joshi@roslin.ed.ac.uk

Keywords:

Single cell RNA-seq

Embryonic stem cells

Transcription control

Gene expression variance

Coefficient of variance

Abstract

Background

Gene expression heterogeneity contributes to development as well as disease progression. Due to technological limitations, most studies to date have focused on differences in mean expression across experimental conditions, rather than differences in gene expression variance. The advent of single cell RNA sequencing has now made it feasible to study gene expression heterogeneity and to characterise genes based on their coefficient of variation.

Methods

We collected single cell gene expression profiles for 32 human and 39 mouse embryonic stem cells and studied correlation between diverse characteristics such as network connectivity and coefficient of variation (CV) across single cells. We further systematically characterised properties unique to High CV genes.

Results

Highly expressed genes tended to have a low CV and were enriched for cell cycle genes. In contrast, High CV genes were co-expressed with other High CV genes, were enriched for bivalent (H3K4me3 and H3K27me3) marked promoters and showed enrichment for response to DNA damage and DNA repair.

Conclusions

Taken together, this analysis demonstrates the divergent characteristics of genes based on their CV. High CV genes tend to form co-expression clusters and they explain bivalency at least in part.

Background

Transcription control is fundamental to mammalian system in defining gene expression programs that establish and maintain specific cell states during development. Any aberration to this process can result into disease phenotype. Microarray technology enables a genome-wide snapshot of the transcription

landscape during development and disease by parallel quantification of large numbers of messenger RNA transcripts from different cell types and tissues(Schulze and Downward, 2001). This technology is widely used for differential gene expression analysis where studies are performed on a pool of hundreds of thousands of cells with an assumption that the variation across multiple samples from a cell population is largely due to experimental noise. Difference between mean values of gene expression is therefore the focus of such analyses and rarely the variability across the samples(Mar et al., 2011).

The breakthroughs in sequencing technology have now made it feasible to generate gene expression data for hundreds of individual cells from a cell population(Pan, 2014) providing new insights into early development(Tang et al., 2010) and differentiation(Shalek et al., 2013). Single cell RNA-seq sequencing is used for characterisation of hidden subpopulations of rare cell types, as closely related cells with the same phenotype can be discriminated to distinguish functionally each subgroup(Buettner et al., 2015). Importantly, the gene expression quantification by single-cell RNA-seq is consistent with the existing gold standards(Wu et al., 2013). The single cell gene expression data is variable between individual cells in contrast to the high concordance across replicates of populations of cells(Shalek et al., 2013). Though part of variation across individual cells is attributed to various confounding factors such as random technical noise mainly due to transcription bursts (Brennecke et al., 2013), protein fluctuations (Karwacki-Neisius et al., 2013) or mRNA fluctuations in response to cell cycle(Singh et al., 2013), there is no doubt about the biological relevance of variation in development(Xue et al., 2013), evolutionary adaptation, and disease(Feinberg and Irizarry, 2010).

Importantly, variation at a single cell level in genetically identical organisms in homogeneous environments indicates its role in generating diversity(Raj et al., 2010). Achieving such diversity is particularly important in the context of stem cells. The pluripotent state is a delicate equilibrium between the ability of self-renewal and differentiation, hence an imbalance (the variation of key pluripotency factors) could lead tipping the scale in favour of differentiation(Karwacki-Neisius et al., 2013). Accordingly, a high concordance was noted between global gene expression variability and heterogeneity of human pluripotency states(Mason et al., 2014). The differences between gene sets at the two ends of the spectrum of variation demonstrated that low variance genes were highly connected in the regulatory networks providing a causal hypothesis for their low variance(Mar et al., 2011). Highly variable genes, on the other hand, are thought to represent elements which fluctuate as the stem cell population moves between self-renewal and differentiation-potential(Mason et al., 2014). We collected single cell RNA sequencing data in human (Streets et al., 2014) and mouse (Yan et al., 2013) embryonic stem cells and identified 'High CV' (CV: Coefficient of Variation) gene sets. The multi-facetted bioinformatic analysis was based on CV enabled systematic characterisation of differences between the stable and variable gene sets.

Methods

Data collection and processing: Single cell RNA-seq data was obtained from Gene Expression Omnibus (GEO) database (Barrett et al., 2013) in fastq format. We downloaded 63 mouse single ES cell RNA-seq data (paired end) (GSE47835, SRP025171) (Streets et al., 2014) and 32 human single ES cell RNA-seq data (single end) (GSE36552, SRP011546) (Yan et al., 2013). After quality control using FastQC 0.11.2, alignment was done with TopHat 2.0.9(Trapnell et al., 2009) using mm10 and

hg38 as reference genomes and the GENCODE(Harrow et al., 2012) annotations (M4 and 22) for mouse and human respectively. Expression values for each single cell were calculated following the Cufflinks 2.2.1(Trapnell et al., 2010) pipeline. The aligned reads were converted to expression values using the cuffquant command. Gene expression values for all single cell libraries were generated using the cuffnorm command with the default library normalization method (geometric). 39 mouse ES cells were selected for final analysis after discarding 24 cells due to low read quality or poor alignment scores.

Biological over technical variation threshold: From the initial normalized FPKM value matrix, we discarded the genes with 35 or more, zero expression values for mouse and 28 or more, zero expression values for human. We calculated the mean FPKM values (mean expression) across all cells for each of the remaining genes. We selected 229 (mESCs) and 217 (hESCs) highly expressed genes (> 150 FPKM is each single cell) as highly confident sets. The remaining genes were sorted according to their mean expression levels and divided in windows of 1,000 genes each (16 windows mouse, 19 windows human). The lowest windows (1,259 genes in mouse, 1,025 genes in human) were comprised of genes with the lowest mean expression levels, hence suffering from high levels of technical variation. We calculated the Pearson correlation coefficient for each pair of highly expressed genes with each gene in each window. For each window, (except the lowest one) we compared the distribution of correlation of all the gene pairs with the distribution of correlation of the lowest window using a t-test. We kept the genes with significantly higher correlation (probability distribution shifted to the right) compared to the lowest window (comparable to random noise). CV was determined as the ration of standard deviation to mean for each gene across single cells.

Transcription factor enrichment: We used data from 49 and 99 ChIP-seq experiments for transcription factors and chromatin remodellers in human and mouse embryonic stem cells respectively (Pooley et al., 2014). We selected peaks in promoter regions (+/- 1kb from the TSS) of the two groups (High CV and Non High CV). For each promoter region, we also counted the total number of factors binding at the region.

miRNA target interactions: Data of miRNA target interactions in ES cells were retrieved from the ESCAPE database (Xu et al., 2013). From 693,552 interactions, we kept only the interactions that their target genes were in our one-to-one orthologs list and divided the number of miRNA interactions per gene in 3 bins (1-50, 51-100, >100).

Protein-Protein interactions: Data of protein-protein interactions were retrieved from the ESCAPE database (Xu et al., 2013). One-to-one orthologs were used to map the genes for each category and for the total list of interactions. The number of proteins interacting with each gene were divided in four bins (1, 2, 3, >3).

Overlap with bivalent and active genes: We overlapped our genes with genes that were classified as bivalent or active (H3K4me3 marked) in human and mouse ES cells using unpublished work from our lab (Mantsoki et al., *accepted*) and studied their differences at the level of CV.

Overlap with CpG islands and TATA box promoters: We calculated the overlap of the promoters of the genes with the CpG island regions as given from the UCSC tracks unmasked CpG islands for hg38 and mm10 (Karolchik et al., 2014). 2,742 murine and 2,010 human TATA-box motif promoters were retrieved from the Eukaryotic Promoter Database(Dreos et al., 2015).

Gene type classification: We calculated the fraction of genes that belonged to a specific gene type (from GENCODE annotation files). We selected only the types of genes with at least 30 genes in all the groups and plotted the CV for each category.

High variation threshold: For the sets of genes that were above the threshold of technical noise we calculated the coefficient of variation (CV) using the standard definition of ratio of the standard deviation to the mean, and divided them in four groups (quartiles) according to their CV. The High variation (High CV) genes were the ones that were falling in the fourth quartile of the CV. The rest of the genes were defined as Non High CV. Gene ontology enrichment was performed using DAVID (Dennis et al., 2003).

Correlation co-expression analysis: We calculated the Pearson correlation coefficient between all the pairs of High CV genes using FPKM values. We randomly permutated the FPKM values between cells for each gene to generate random data. The correlation distributions of High CV genes were significantly different (Wilcoxon test) than the random ones and we investigated their co-expression patterns by hierarchical clustering (flashClust package in R) visualised with heatmaps (heatmap.2 in R).

Conservation analysis: 17,009 one-to-one orthologs from ensembl BioMart (Guberman et al., 2011) were used to calculate CV values in each species. After intersecting the orthologs with the 4,000 genes (for both mouse and human) we end up with a gene set containing 2,363 orthologous genes.

Topological associated domains: A lists of topological associated domains (TADs) for mouse and human ES cells (Dixon et al., 2012) was used to calculate the number of genes per TAD for the High CV and Non High CV genes in our analysis.

Bulk expression data: For the bulk RNA analysis we used 3 biological replicates of Microarray data from mouse ES cells (GSM1326660-2) (Zhang et al., 2014) and 4 biological replicates of RNA-seq data from hESCs (GSE33480) (Djebali et al., 2012).

Sequence conservation: The sequence conservation scores where obtained from PhyloP100way (Human) and PhyloP60way (Mouse) tracks available at UCSC.

Results

Correlation based approach to identify genes with significant biological variation in mammalian single embryonic stem cell RNA-seq data

To study the gene expression variability across individual cells, we collected RNA sequencing data for 32 human and 39 mouse single ES cells. After normalising the data across cells, we calculated FPKM values for 43,345 mouse and 60,468 human GENCODE (Harrow et al., 2012) genes in each single cell. Single cell sequencing data suffers from low genome coverage and high amplification bias. These biases contribute to technical variation (noise) which hinders capturing biological variation across individual single cells. To distinguish the genes with significantly higher biological variation over technical variation, we developed a correlation-based approach. As highly expressed genes tend to have lower technical noise, we selected top 229 (mouse) and 217 (human) highly expressed genes (see Methods) across single cells. We then binned the genes based on their mean expression level. We calculated the correlation of genes in each bin with the highly expressed genes. We noted that technical noise was inversely related to the mean expression of gene sets i.e. higher the gene expression, lower the technical noise. We selected a threshold on expression value where the correlation with highly expressed genes was statistically significant over correlation with gene sets with technical noise (see Methods). This procedure resulted in selection of 4229 genes over 2.9 mean expression threshold (log(FPKM+1)) in murine ES cells (Figures 1A and S1) and 4217 genes over log

mean expression threshold of 3.1 in human ES cells (Figures 1B and S2) with significantly higher biological noise than technical noise.

Gene expression variability was negatively correlated with the mean expression level i.e. highly expressed genes had low CV while lowly expressed genes spanned a wide spectrum on CV range (Figures 1A and 1B). The functional enrichment of low CV genes resulted in enrichment for cell cycle functional category specifically the 'M phase' of mitotic cell cycle for both human and mouse ES cells. We further calculated the functional enrichment for highly expressed genes irrespective of CV values. They were also enriched for cell cycle functional category in both human and mouse ES cells. We therefore inferred that highly expressed genes tend to have low CV and are involved in cellular functions such as cell cycle.

We further checked if different gene categories provided by GENCODE (Harrow et al., 2012) demonstrate variability comparable to protein coding genes (Figure 1C and D). The lincRNAs had higher CV values in both human (t-test, P-value < 0.01) and mouse ES cells (t-test, P-value < 0.05). An overwhelming fraction of murine processed pseudogenes had low CV (t-test, P-value < 0.05). In contrast, a significant fraction of human processed pseudogenes had CV higher than protein-coding genes (t-test, P-value < 0.001). Processed transcripts and antisense transcripts on the other hand show no significant difference, possibly due to low sample numbers.

Genes occupied by many transcription factors have a lower CV

In order to study the level of transcription control among three groups of promoters, we calculated the number of factors binding at each promoter using ChIP sequencing compendia for transcription and epigenetic factors in human and ES cells(Pooley et al., 2014). The mean CV for genes bound by less than 10 factors was significantly higher than the mean CV for genes bound by more than 10 factors in both human (t-

test, P-value < 0.001) and mouse (t-test, P-value < 0.001) ES cells (Figure 2A and B). This result was consistent when average binding of individual factors was tested as well i.e. genes more likely to be bound by more factors tended to have low CV. We obtained the number of putative binding sites of transcription factors in gene promoters from UCSC. Again, number of putative binding sites varied inversely with the CV value (Figure S3).

To test the regulation at post-transcriptional level, we collected putative miRNA targets predicted by four miRNA prediction methods(Xu et al., 2013). Unlike TF targets, there was no bias towards the number of miRNA targets with respect to their mean CV, either in human or mouse ES cells (Figure 2C and D).

Finally we collected known protein-protein interactions (PPI) in mouse and human ES cells(Xu et al., 2013) and calculated the number of known interacting partners for each of the genes. Similarly to miRNA targets, there was no statistically significant difference between the mean CV values based on the number of interacting partners at protein level in either human or mouse ES cells (Figure 2E and F).

High expression variability genes correlate with DNA repair and bivalency

The activity of signalling pathways such as TGF-β-related signalling pathways are thought to prime cells for differentiation contributing to the heterogeneity between cells in ES cells (Galvin-Burgess et al., 2013). The CV value did not distinguish any particular signalling pathway. The differences in micro-environments sensed by the signalling pathway can manifest in large expression changes of its downstream target genes. We therefore tested whether transcription factor and chromatin remodeller binding prefers or avoids gene promoters based on their CV measure using the ChIP sequencing data compendium for 49 and 99 factors in mouse and human ES cells

respectively(Pooley et al., 2014). Unsurprisingly, many promoter specific factors such as E2F1, TAF1, and YY1 did not show any bias for the CV. High CV genes in mouse ES cells showed an exclusive binding preference of the following four factors: NCOA3 (Hypergeometric test, P-value < 0.0001), p300 (Hypergeometric test, P-value < 0.0001), MCAF1 (Hypergeometric test, P-value < 0.01) and p53(Hypergeometric test, P-value < 0.05).

NCOA3 is a nuclear receptor activator with a histone acetyltransferase activity, recruiting the chromatin modifying proteins p300, CARM1 and CBP at the *Nanog* locus (Wu et al., 2012). NCOA3 is thought to be critical for both the induction and maintenance of pluripotency, acting as an essential Esrrb coactivator (Percharde et al., 2012). ESRRB is downstream of NANOG which is a direct target of TGF-β mediated SMAD signalling(Xu et al., 2008). NANOG targets did not show any bias with respect to CV.

MCAF1 is a nuclear protein associated with heterochromatin, shown to colocalize with SETDB1 in PML bodies (Sasai et al., 2013). PML is a protein involved in the senescence pathway through the p53 signalling, and its overexpression leads to premature senescence (Pearson et al., 2000). p53 is a sequence specific transcription factor with tumour suppressor activity, regulating cell cycle arrest, apoptosis, senescence and stem cell differentiation, acting as an activator or suppressor of its downstream targets (Vousden and Prives, 2009). Upon DNA damage, p53 activates differentiation associated genes and represses self-renewal genes, affecting the status of ES cells (Li et al., 2012).

Accordingly, high CV genes showed enrichment for biological processes such as cellular response to stress (adjusted P-value < 10^{-4}), response to DNA damage

stimulus (adjusted P-value < 10^{-3}) and DNA repair (adjusted P-value < 10^{-3}) in both murine and human ES cells.

The genes overlapping with bivalent promoters had statistically significant higher CV values than the ones overlapping with the active promoters (presence of H3K4me3 and absence of H3K27me3 modifications) in both human (Hypergeometric test, P-value < 0.001) and mouse (Hypergeometric test, P-value < 0.001) ES cells (Figure 3A and 3B). Genes with high CV showed a weak functional enrichment for embryonic development and transcription control; the functional categories associated with bivalent genes(Bernstein et al., 2006).

As specific promoter structures such as presence of TATA boxes have been previously associated with genes with highly fluctuating single-cell levels within populations(Choi and Kim, 2009), we calculated TATA and CpG island fraction for all human and mouse promoters (-/+ 1Kb from TSS). The CpG-rich promoters showed lower CV values than the CpG-poor promoters and the difference was statistically significant in both human and mouse ES cells (t-test P-value<0.001) (Figure 3C and D). Unlike CpG promoters, TATA box promoters could not be distinguished based on the CV value (Figure 3E and F).

High CV genes form dense highly co-expressed clusters

In order to study the characteristics of genes with high variability, we defined genes with CV value greater than 0.92 (3rd quartile value) as High CV in mouse (Figure 4A) and genes with CV value greater than 1.45 (3rd quartile value) in human ES cells (Figure 4B). We then checked whether the expression of High CV genes varies concordantly across single cells by calculating Pearson's correlation coefficient between all pairs of High CV genes. A subset of High CV genes were significantly

more correlated with each other compared to expected from a random permutation (Figures 4C (mouse) and 4D (human)).

The highly correlated network (Pearson's correlation coefficient > 0.95) of High CV genes grouped them mainly into only few tightly co-expressed clusters in both human and mouse ES cells (Figures S4 and S5). Interestingly, the genes in each cluster were highly expressed only in one individual cell (Figure 4E (mouse) and 4F (human)). We firstly confirmed that these single cells (e.g. single cell 24 and 26 in humans) did not suffer from poor technical quality of samples (Figure S6). We also removed these two cells and redefined the High CV gene set (Figure S7) to find a similar result. This assured that the significant co-expression among High CV genes is not an artefact of few aberrant single cells.

The co-expressed genes derived from large-scale analyses of mammalian expression data have demonstrated that neighbouring genes tend to have similar expression profiles(Lercher et al., 2002). As high CV genes formed tight co-expression clusters, we checked whether they tend to be in gene neighbourhoods with each other compared to other genes. We did not observe any tendency of genes clustering based on CV value. We also checked whether there was any bias towards similar CV genes co-existing in topological associated domains (TADS) inferred from Hi-C chromatin capture data in human and mouse ES cells (Dixon et al., 2012). There was no bias towards associating similar CV value genes with same TADS. Also, tightly co-expressed High CV genes in each cluster were not specifically enriched for any biological process nor primed for specific lineage.

CV values are conserved across species

In order to check whether the CV values are conserved between bulk and single cell experiments, we obtained gene expression values for bulk RNA in human and mouse

ES cells. The CV values of genes from single cells and bulk RNA showed no correlation in both human (Pearson's correlation coefficient r=0.09) and mouse (Pearson's correlation coefficient r=0.06) ES cells (Figure 5A and B).

To test whether gene expression variability from single and bulk RNA-seq is conserved across species, we collected one-to-one orthologs between human and mouse (Guberman et al., 2011). The gene expression tends to be conserved across species for single (Pearson's correlation coefficient r=0.23) (Figure 5C) i.e. orthologs of genes with lower CV in mouse are more likely to have lower expression variance across human single ES cells and vice versa. We confirmed that the distribution of CV values for orthologous genes in mouse was not significantly different from mouse-specific genes (Figure 5D). We further checked whether the expression conservation goes hand-in-hand with the conservation at the sequence level. Indeed, sequence conservation showed a negative correlation with the CV values in both human and mouse ES cells in their 5'UTR, their 3'UTR and their exons (Figure 5E and F). Thus tight regulation of gene expression level is a feature that appears to be conserved and selected during evolution.

Conclusion and Discussion

Single cell RNA-seq data holds a great promise for studying variability across individual cells with the hindrance of large technical noise inherent to these data. Though availability of data from a limited number of cells (32 in human, 39 in mouse) could influence the results, it has been recently shown that 30 cells is the lower limit of sample size to sufficiently converge to the complexity of large cell populations (Marinov et al., 2014). We used a correlation based approach to define a set of genes with biological variation significantly higher than technical variation across single cells. We then studied the characteristics of expression variability for 4,217 genes in

human and 4,229 genes in mouse single ES cells, where the estimated biological variability was significantly greater than the technical variability. We noted that highly expressed genes tended to have lower CV (Figure 1A & B). Since ES cells are not synchronized in their cell cycle and can belong to different development stages, we specifically looked whether genes with high CV were developmental stage specific or involved in specific function, but did not find a strong evidence for it. High CV genes form co-expression clusters. Tightly co-expressed High CV genes in each cluster were highly expressed only in one or a few single cell(s) and genes in each cluster were not specifically enriched for any biological process. This fits with the notion of pluripotent cells to alternate between different transient and reversible cell states where transient states do not show any functional bias or lineage priming. High CV genes showed enrichment for response to DNA damage and DNA repair and were exclusively bound by regulators of DNA damage and senescence pathways like MCAF1 and p53. They also showed significant overlap with bivalent genes in human and mouse ES cells. This confirms that at least a subset of bivalent genes can indeed be attributed to heterogeneity in ES cells.

Though many characteristics of CV genes are conserved across species, there are some differences. Interestingly the vast majority of murine processed pseudogenes have lower CV than protein-coding genes while human processed pseudogenes have higher CV than protein-coding genes. Processed pseudogenes have recently been demonstrated to play a regulatory role by competing with other genes for the binding of small RNAs(Poliseno et al., 2010). This potential species specific regulatory aspect needs to be explored in detail.

Taken together, genes with lower CV tend to be highly expressed, tightly regulated at transcriptional level as they are likely to be central to many cellular processes. High

CV genes, on the other hand, are highly expressed only in individual single cells which possibly partly explains the bivalent genes (with both active and inactive chromatin status) observed in bulk studies.

Competing interests

The authors declare no completing interests.

Authors' contributions

A.M. collected the data, performed the analysis and helped write the manuscript, G.D. helped perform the analysis, and A.J. conceived the idea, supervised the project and wrote the manuscript.

Acknowledgements

AJ is a Chancellor's fellow at the Roslin Institute, University of Edinburgh. A.J. lab is supported by Biotechnology and Biological Sciences Research Council (BBSRC).

References

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S., Soboleva, A., 2013. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 41, D991–995. doi:10.1093/nar/gks1193

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S.L., Lander, E.S., 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125, 315–326. doi:10.1016/j.cell.2006.02.041

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., Heisler, M.G., 2013. Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods 10, 1093–1095. doi:10.1038/nmeth.2645

Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., Stegle, O., 2015. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol. 33, 155–160. doi:10.1038/nbt.3102

Choi, J.K., Kim, Y.-J., 2009. Intrinsic variability of gene expression encoded in nucleosome positioning sequences. Nat. Genet. 41, 498–503. doi:10.1038/ng.319

Dennis, G., Sherman, B.T., Hosack, D. a, Yang, J., Gao, W., Lane, H.C., Lempicki, R.

a, 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 4, P3.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., Ren, B., 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380. doi:10.1038/nature11082

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G.K., Khatun, J., Williams, B.A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Bar, N.S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M.J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O.J., Park, E., Persaud, K., Preall, J.B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S.E., Hannon, G., Giddings, M.C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., Gingeras, T.R., 2012. Landscape of transcription in human cells. Nature 489, 101–8. doi:10.1038/nature11233

Dreos, R., Ambrosini, G., Périer, R.C., Bucher, P., 2015. The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. Nucleic Acids Res. 43, D92–96. doi:10.1093/nar/gku1111

Feinberg, A.P., Irizarry, R.A., 2010. Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proc. Natl. Acad. Sci. U. S. A. 107 Suppl , 1757–1764. doi:10.1073/pnas.0906183107

Galvin-Burgess, K.E., Travis, E.D., Pierson, K.E., Vivian, J.L., 2013. TGF-β-superfamily signaling regulates embryonic stem cell heterogeneity: self-renewal as a dynamic and regulated equilibrium. Stem Cells 31, 48–58. doi:10.1002/stem.1252

Guberman, J.M., Ai, J., Arnaiz, O., Baran, J., Blake, A., Baldock, R., Chelala, C., Croft, D., Cros, A., Cutts, R.J., Di Génova, A., Forbes, S., Fujisawa, T., Gadaleta, E., Goodstein, D.M., Gundem, G., Haggarty, B., Haider, S., Hall, M., Harris, T., Haw, R., Hu, S., Hubbard, S., Hsu, J., Iyer, V., Jones, P., Katayama, T., Kinsella, R., Kong, L., Lawson, D., Liang, Y., Lopez-Bigas, N., Luo, J., Lush, M., Mason, J., Moreews, F., Ndegwa, N., Oakley, D., Perez-Llamas, C., Primig, M., Rivkin, E., Rosanoff, S., Shepherd, R., Simon, R., Skarnes, B., Smedley, D., Sperling, L., Spooner, W., Stevenson, P., Stone, K., Teague, J.,

Wang, J., Wang, J., Whitty, B., Wong, D.T., Wong-Erasmus, M., Yao, L., Youens-Clark, K., Yung, C., Zhang, J., Kasprzyk, A., 2011. BioMart Central Portal: an open database network for the biological community. Database (Oxford). 2011, bar041. doi:10.1093/database/bar041

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J.M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R., Hubbard, T.J., 2012. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760–1774. doi:10.1101/gr.135350.111

Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., Harte, R.A., Heitner, S., Hinrichs, A.S., Learned, K., Lee, B.T., Li, C.H., Raney, B.J., Rhead, B., Rosenbloom, K.R., Sloan, C.A., Speir, M.L., Zweig, A.S., Haussler, D., Kuhn, R.M., Kent, W.J., 2014. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 42, D764–70. doi:10.1093/nar/gkt1168

Karwacki-Neisius, V., Göke, J., Osorno, R., Halbritter, F., Ng, J.H., Weiße, A.Y., Wong, F.C.K., Gagliardi, A., Mullin, N.P., Festuccia, N., Colby, D., Tomlinson, S.R., Ng, H.-H., Chambers, I., 2013. Reduced Oct4 expression directs a robust pluripotent state with distinct signaling activity and increased enhancer occupancy by Oct4 and Nanog. Cell Stem Cell 12, 531–545. doi:10.1016/j.stem.2013.04.023

Lercher, M.J., Urrutia, A.O., Hurst, L.D., 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat. Genet. 31, 180–183. doi:10.1038/ng887

Li, M., He, Y., Dubois, W., Wu, X., Shi, J., Huang, J., 2012. Distinct regulatory mechanisms and functions for p53-activated and p53-repressed DNA damage response genes in embryonic stem cells. Mol. Cell 46, 30–42. doi:10.1016/j.molcel.2012.01.020

Mar, J.C., Matigian, N.A., Mackay-Sim, A., Mellick, G.D., Sue, C.M., Silburn, P.A., McGrath, J.J., Quackenbush, J., Wells, C.A., 2011. Variance of Gene Expression Identifies Altered Network Constraints in Neurological Disease. PLoS Genet. 7, e1002207. doi:10.1371/journal.pgen.1002207

Marinov, G.K., Williams, B.A., McCue, K., Schroth, G.P., Gertz, J., Myers, R.M., Wold, B.J., 2014. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res. 24, 496–510.

doi:10.1101/gr.161034.113

Mason, E.A., Mar, J.C., Laslett, A.L., Pera, M.F., Quackenbush, J., Wolvetang, E., Wells, C.A., 2014. Gene Expression Variability as a Unifying Element of the Pluripotency Network. Stem Cell Reports 3, 365–377. doi:10.1016/j.stemcr.2014.06.008

Pan, X., 2014. Single Cell Analysis: From Technology to Biology and Medicine. Single Cell Biol. 3. doi:10.4172/2168-9431.1000106

Pearson, M., Carbone, R., Sebastiani, C., Cioce, M., Fagioli, M., Saito, S., Higashimoto, Y., Appella, E., Minucci, S., Pandolfi, P.P., Pelicci, P.G., 2000. PML regulates p53 acetylation and premature senescence induced by oncogenic Ras. Nature 406, 207–10. doi:10.1038/35018127

Percharde, M., Lavial, F., Ng, J.-H., Kumar, V., Tomaz, R.A., Martin, N., Yeo, J.-C., Gil, J., Prabhakar, S., Ng, H.-H., Parker, M.G., Azuara, V., 2012. Ncoa3 functions as an essential Esrrb coactivator to sustain embryonic stem cell self-renewal and reprogramming. Genes Dev. 26, 2286–98. doi:10.1101/gad.195545.112

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W.J., Pandolfi, P.P., 2010. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 465, 1033–1038. doi:10.1038/nature09144

Pooley, C., Ruau, D., Lombard, P., Gottgens, B., Joshi, A., 2014. TRES predicts transcription control in embryonic stem cells. Bioinformatics 30, 2983–2985. doi:10.1093/bioinformatics/btu399

Raj, A., Rifkin, S.A., Andersen, E., van Oudenaarden, A., 2010. Variability in gene expression underlies incomplete penetrance. Nature 463, 913–918. doi:10.1038/nature08781

Sasai, N., Saitoh, N., Saitoh, H., Nakao, M., 2013. The transcriptional cofactor MCAF1/ATF7IP is involved in histone gene expression and cellular senescence. PLoS One 8, e68478. doi:10.1371/journal.pone.0068478

Schulze, A., Downward, J., 2001. Navigating gene expression using microarrays--a technology review. Nat. Cell Biol. 3, E190–195. doi:10.1038/35087138

Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., Trombetta, J.J., Gennert, D., Gnirke, A., Goren, A., Hacohen, N., Levin, J.Z., Park, H., Regev, A., 2013. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature 498, 236–240. doi:10.1038/nature12172

Singh, A.M., Chappell, J., Trost, R., Lin, L., Wang, T., Tang, J., Matlock, B.K., Weller, K.P., Wu, H., Zhao, S., Jin, P., Dalton, S., 2013. Cell-cycle control of

developmentally regulated transcription factors accounts for heterogeneity in human pluripotent cells. Stem Cell Reports 1, 532–544. doi:10.1016/j.stemcr.2013.10.009

Streets, A.M., Zhang, X., Cao, C., Pang, Y., Wu, X., Xiong, L., Yang, L., Fu, Y., Zhao, L., Tang, F., Huang, Y., 2014. Microfluidic single-cell whole-transcriptome sequencing. Proc. Natl. Acad. Sci. U. S. A. 111, 7048–53. doi:10.1073/pnas.1402030111

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., Lao, K., Surani, M.A., 2010. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell 6, 468–478. doi:10.1016/j.stem.2010.03.015

Trapnell, C., Pachter, L., Salzberg, S.L., 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111. doi:10.1093/bioinformatics/btp120

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511–515. doi:10.1038/nbt.1621

Vousden, K.H., Prives, C., 2009. Blinded by the Light: The Growing Complexity of p53. Cell 137, 413–31. doi:10.1016/j.cell.2009.04.037

Wu, A.R., Neff, N.F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M.E., Mburu, F.M., Mantalas, G.L., Sim, S., Clarke, M.F., Quake, S.R., 2013. Quantitative assessment of single-cell RNA-sequencing methods. Nat. Methods 11, 41–46. doi:10.1038/nmeth.2694

Wu, Z., Yang, M., Liu, H., Guo, H., Wang, Y., Cheng, H., Chen, L., 2012. Role of nuclear receptor coactivator 3 (Ncoa3) in pluripotency maintenance. J. Biol. Chem. 287, 38295–304. doi:10.1074/jbc.M112.373092

Xu, H., Baroukh, C., Dannenfelser, R., Chen, E.Y., Tan, C.M., Kou, Y., Kim, Y.E., Lemischka, I.R., Ma'ayan, A., 2013. ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. Database (Oxford). 2013, bat045. doi:10.1093/database/bat045

Xu, R.-H., Sampsell-Barron, T.L., Gu, F., Root, S., Peck, R.M., Pan, G., Yu, J., Antosiewicz-Bourget, J., Tian, S., Stewart, R., Thomson, J.A., 2008. NANOG is a direct target of TGFbeta/activin-mediated SMAD signaling in human ESCs. Cell Stem Cell 3, 196–206. doi:10.1016/j.stem.2008.07.001

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y.E., Liu, J., Horvath, S., Fan, G., 2013. Genetic programs in human and

mouse early embryos revealed by single-cell RNA sequencing. Nature 500, 593–597. doi:10.1038/nature12364

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., Huang, J., Li, M., Wu, X., Wen, L., Lao, K., Li, R., Qiao, J., Tang, F., 2013. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat. Struct. Mol. Biol. 20, 1131–1139. doi:10.1038/nsmb.2660

Zhang, Y., Xie, S., Zhou, Y., Xie, Y., Liu, P., Sun, M., Xiao, H., Jin, Y., Sun, X., Chen, Z., Huang, Q., Chen, S., 2014. H3K36 histone methyltransferase Setd2 is required for murine embryonic stem cell differentiation toward endoderm. Cell Rep. 8, 1989–2002. doi:10.1016/j.celrep.2014.08.031

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S., Soboleva, A., 2013. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 41, D991–995. doi:10.1093/nar/gks1193

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S.L., Lander, E.S., 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125, 315–326. doi:10.1016/j.cell.2006.02.041

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., Heisler, M.G., 2013. Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods 10, 1093–1095. doi:10.1038/nmeth.2645

Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., Stegle, O., 2015. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol. 33, 155–160. doi:10.1038/nbt.3102

Choi, J.K., Kim, Y.-J., 2009. Intrinsic variability of gene expression encoded in nucleosome positioning sequences. Nat. Genet. 41, 498–503. doi:10.1038/ng.319

Dennis, G., Sherman, B.T., Hosack, D. a, Yang, J., Gao, W., Lane, H.C., Lempicki, R. a, 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 4, P3.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., Ren, B., 2012. Topological domains in mammalian genomes identified by analysis of

chromatin interactions. Nature 485, 376–380. doi:10.1038/nature11082

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G.K., Khatun, J., Williams, B.A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Bar, N.S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M.J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O.J., Park, E., Persaud, K., Preall, J.B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S.E., Hannon, G., Giddings, M.C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., Gingeras, T.R., 2012. Landscape of transcription in human cells. Nature 489, 101–8. doi:10.1038/nature11233

Dreos, R., Ambrosini, G., Périer, R.C., Bucher, P., 2015. The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. Nucleic Acids Res. 43, D92–96. doi:10.1093/nar/gku1111

Feinberg, A.P., Irizarry, R.A., 2010. Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proc. Natl. Acad. Sci. U. S. A. 107 Suppl , 1757–1764. doi:10.1073/pnas.0906183107

Galvin-Burgess, K.E., Travis, E.D., Pierson, K.E., Vivian, J.L., 2013. TGF-β-superfamily signaling regulates embryonic stem cell heterogeneity: self-renewal as a dynamic and regulated equilibrium. Stem Cells 31, 48–58. doi:10.1002/stem.1252

Guberman, J.M., Ai, J., Arnaiz, O., Baran, J., Blake, A., Baldock, R., Chelala, C., Croft, D., Cros, A., Cutts, R.J., Di Génova, A., Forbes, S., Fujisawa, T., Gadaleta, E., Goodstein, D.M., Gundem, G., Haggarty, B., Haider, S., Hall, M., Harris, T., Haw, R., Hu, S., Hubbard, S., Hsu, J., Iyer, V., Jones, P., Katayama, T., Kinsella, R., Kong, L., Lawson, D., Liang, Y., Lopez-Bigas, N., Luo, J., Lush, M., Mason, J., Moreews, F., Ndegwa, N., Oakley, D., Perez-Llamas, C., Primig, M., Rivkin, E., Rosanoff, S., Shepherd, R., Simon, R., Skarnes, B., Smedley, D., Sperling, L., Spooner, W., Stevenson, P., Stone, K., Teague, J., Wang, J., Wang, J., Whitty, B., Wong, D.T., Wong-Erasmus, M., Yao, L., Youens-Clark, K., Yung, C., Zhang, J., Kasprzyk, A., 2011. BioMart Central Portal: an open database network for the biological community. Database (Oxford). 2011, bar041. doi:10.1093/database/bar041

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J.M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R., Hubbard, T.J., 2012. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760–1774. doi:10.1101/gr.135350.111

Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., Harte, R.A., Heitner, S., Hinrichs, A.S., Learned, K., Lee, B.T., Li, C.H., Raney, B.J., Rhead, B., Rosenbloom, K.R., Sloan, C.A., Speir, M.L., Zweig, A.S., Haussler, D., Kuhn, R.M., Kent, W.J., 2014. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 42, D764–70. doi:10.1093/nar/gkt1168

Karwacki-Neisius, V., Göke, J., Osorno, R., Halbritter, F., Ng, J.H., Weiße, A.Y., Wong, F.C.K., Gagliardi, A., Mullin, N.P., Festuccia, N., Colby, D., Tomlinson, S.R., Ng, H.-H., Chambers, I., 2013. Reduced Oct4 expression directs a robust pluripotent state with distinct signaling activity and increased enhancer occupancy by Oct4 and Nanog. Cell Stem Cell 12, 531–545. doi:10.1016/j.stem.2013.04.023

Lercher, M.J., Urrutia, A.O., Hurst, L.D., 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat. Genet. 31, 180–183. doi:10.1038/ng887

Li, M., He, Y., Dubois, W., Wu, X., Shi, J., Huang, J., 2012. Distinct regulatory mechanisms and functions for p53-activated and p53-repressed DNA damage response genes in embryonic stem cells. Mol. Cell 46, 30–42. doi:10.1016/j.molcel.2012.01.020

Mar, J.C., Matigian, N.A., Mackay-Sim, A., Mellick, G.D., Sue, C.M., Silburn, P.A., McGrath, J.J., Quackenbush, J., Wells, C.A., 2011. Variance of Gene Expression Identifies Altered Network Constraints in Neurological Disease. PLoS Genet. 7, e1002207. doi:10.1371/journal.pgen.1002207

Marinov, G.K., Williams, B.A., McCue, K., Schroth, G.P., Gertz, J., Myers, R.M., Wold, B.J., 2014. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res. 24, 496–510. doi:10.1101/gr.161034.113

Mason, E.A., Mar, J.C., Laslett, A.L., Pera, M.F., Quackenbush, J., Wolvetang, E., Wells, C.A., 2014. Gene Expression Variability as a Unifying Element of the Pluripotency Network. Stem Cell Reports 3, 365–377.

doi:10.1016/j.stemcr.2014.06.008

Pan, X., 2014. Single Cell Analysis: From Technology to Biology and Medicine. Single Cell Biol. 3. doi:10.4172/2168-9431.1000106

Pearson, M., Carbone, R., Sebastiani, C., Cioce, M., Fagioli, M., Saito, S., Higashimoto, Y., Appella, E., Minucci, S., Pandolfi, P.P., Pelicci, P.G., 2000. PML regulates p53 acetylation and premature senescence induced by oncogenic Ras. Nature 406, 207–10. doi:10.1038/35018127

Percharde, M., Lavial, F., Ng, J.-H., Kumar, V., Tomaz, R.A., Martin, N., Yeo, J.-C., Gil, J., Prabhakar, S., Ng, H.-H., Parker, M.G., Azuara, V., 2012. Ncoa3 functions as an essential Esrrb coactivator to sustain embryonic stem cell self-renewal and reprogramming. Genes Dev. 26, 2286–98. doi:10.1101/gad.195545.112

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W.J., Pandolfi, P.P., 2010. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 465, 1033–1038. doi:10.1038/nature09144

Pooley, C., Ruau, D., Lombard, P., Gottgens, B., Joshi, A., 2014. TRES predicts transcription control in embryonic stem cells. Bioinformatics 30, 2983–2985. doi:10.1093/bioinformatics/btu399

Raj, A., Rifkin, S.A., Andersen, E., van Oudenaarden, A., 2010. Variability in gene expression underlies incomplete penetrance. Nature 463, 913–918. doi:10.1038/nature08781

Sasai, N., Saitoh, N., Saitoh, H., Nakao, M., 2013. The transcriptional cofactor MCAF1/ATF7IP is involved in histone gene expression and cellular senescence. PLoS One 8, e68478. doi:10.1371/journal.pone.0068478

Schulze, A., Downward, J., 2001. Navigating gene expression using microarrays--a technology review. Nat. Cell Biol. 3, E190–195. doi:10.1038/35087138

Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., Trombetta, J.J., Gennert, D., Gnirke, A., Goren, A., Hacohen, N., Levin, J.Z., Park, H., Regev, A., 2013. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature 498, 236–240. doi:10.1038/nature12172

Singh, A.M., Chappell, J., Trost, R., Lin, L., Wang, T., Tang, J., Matlock, B.K., Weller, K.P., Wu, H., Zhao, S., Jin, P., Dalton, S., 2013. Cell-cycle control of developmentally regulated transcription factors accounts for heterogeneity in human pluripotent cells. Stem Cell Reports 1, 532–544. doi:10.1016/j.stemcr.2013.10.009

Streets, A.M., Zhang, X., Cao, C., Pang, Y., Wu, X., Xiong, L., Yang, L., Fu, Y.,

Zhao, L., Tang, F., Huang, Y., 2014. Microfluidic single-cell whole-transcriptome sequencing. Proc. Natl. Acad. Sci. U. S. A. 111, 7048–53. doi:10.1073/pnas.1402030111

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., Lao, K., Surani, M.A., 2010. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell 6, 468–478. doi:10.1016/j.stem.2010.03.015

Trapnell, C., Pachter, L., Salzberg, S.L., 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111. doi:10.1093/bioinformatics/btp120

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511–515. doi:10.1038/nbt.1621

Vousden, K.H., Prives, C., 2009. Blinded by the Light: The Growing Complexity of p53. Cell 137, 413–31. doi:10.1016/j.cell.2009.04.037

Wu, A.R., Neff, N.F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M.E., Mburu, F.M., Mantalas, G.L., Sim, S., Clarke, M.F., Quake, S.R., 2013. Quantitative assessment of single-cell RNA-sequencing methods. Nat. Methods 11, 41–46. doi:10.1038/nmeth.2694

Wu, Z., Yang, M., Liu, H., Guo, H., Wang, Y., Cheng, H., Chen, L., 2012. Role of nuclear receptor coactivator 3 (Ncoa3) in pluripotency maintenance. J. Biol. Chem. 287, 38295–304. doi:10.1074/jbc.M112.373092

Xu, H., Baroukh, C., Dannenfelser, R., Chen, E.Y., Tan, C.M., Kou, Y., Kim, Y.E., Lemischka, I.R., Ma'ayan, A., 2013. ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. Database (Oxford). 2013, bat045. doi:10.1093/database/bat045

Xu, R.-H., Sampsell-Barron, T.L., Gu, F., Root, S., Peck, R.M., Pan, G., Yu, J., Antosiewicz-Bourget, J., Tian, S., Stewart, R., Thomson, J.A., 2008. NANOG is a direct target of TGFbeta/activin-mediated SMAD signaling in human ESCs. Cell Stem Cell 3, 196–206. doi:10.1016/j.stem.2008.07.001

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y.E., Liu, J., Horvath, S., Fan, G., 2013. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature 500, 593–597. doi:10.1038/nature12364

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., Huang, J., Li, M., Wu, X., Wen, L., Lao, K., Li, R., Qiao, J., Tang, F.,

2013. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat. Struct. Mol. Biol. 20, 1131–1139. doi:10.1038/nsmb.2660

Zhang, Y., Xie, S., Zhou, Y., Xie, Y., Liu, P., Sun, M., Xiao, H., Jin, Y., Sun, X., Chen, Z., Huang, Q., Chen, S., 2014. H3K36 histone methyltransferase Setd2 is required for murine embryonic stem cell differentiation toward endoderm. Cell Rep. 8, 1989–2002. doi:10.1016/j.celrep.2014.08.031

Figures

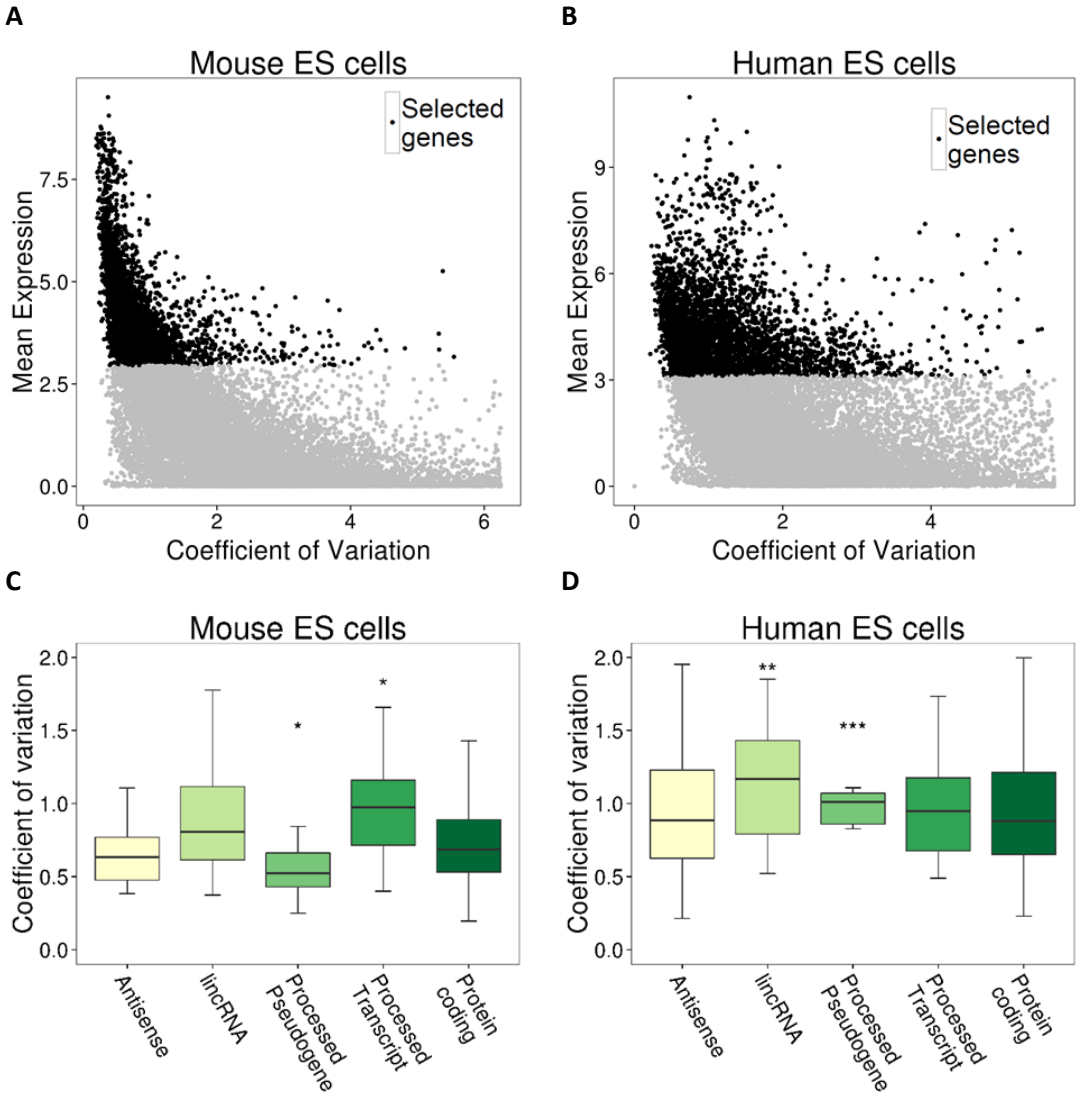


Figure 1. Correlation based approach for the identification of genes above the threshold of technical variation (**A, B**) Scatterplots showing genes according to their mean expression (log (mean FPKM+1)) and coefficient of variation in Mouse and Human ES cells. The genes highlighted in black were chosen for the analysis, since they were more correlated with the highly expressed genes. (**C, D**) Gene types in Mouse and Human ES cells and their respective CV levels (shown only the genes types that were found in 30 genes or more).

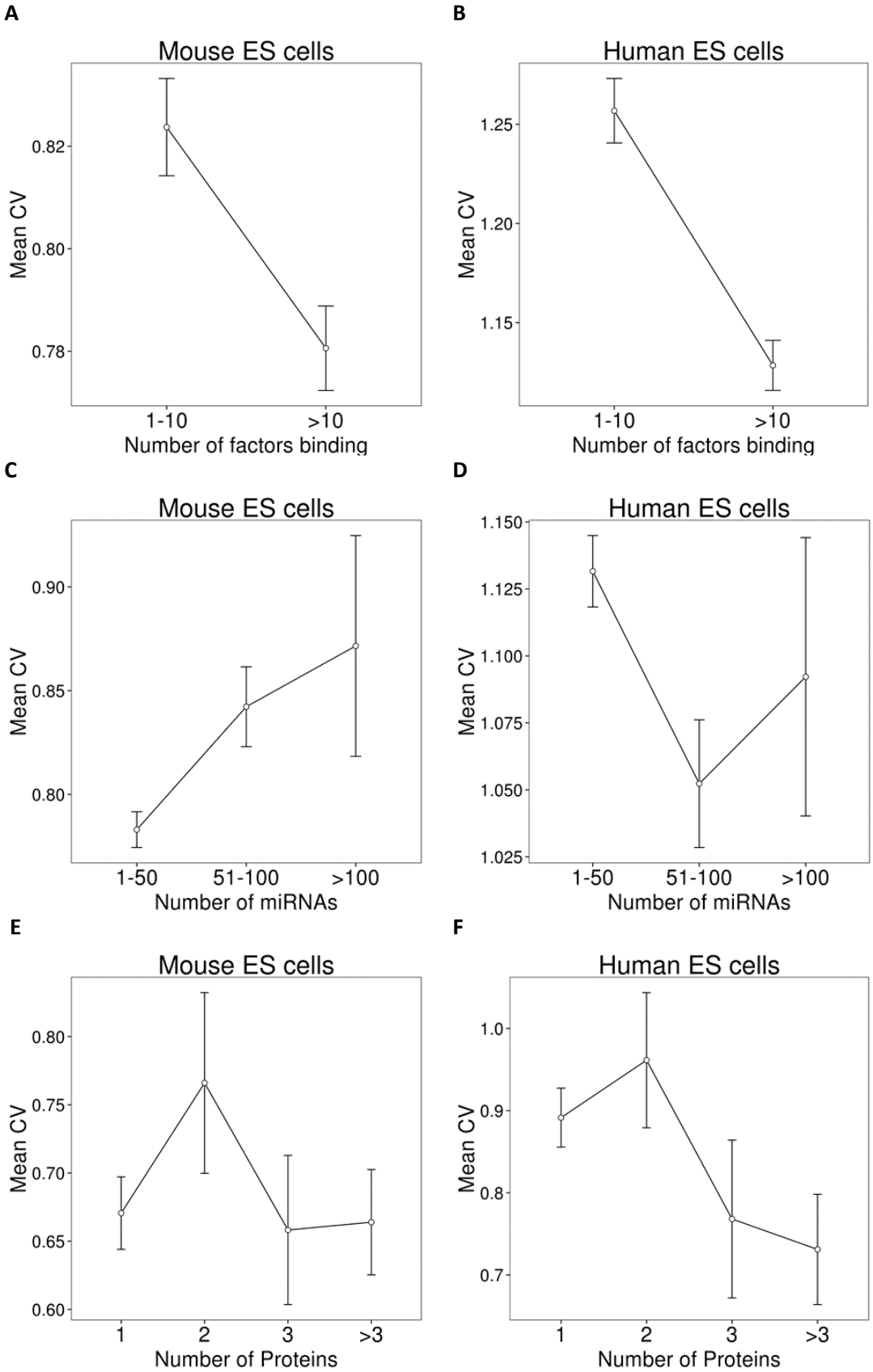


Figure 2. Mean CV levels according to quantification of transcription factors, miRNA targets and protein-protein interactions. (**A, B**) Transcription and epigenetic factor occupancy (number of factors binding) at the promoters of genes is inversely correlated with their Mean CV in Mouse (99 ChIP-seq TFs) and Human (49 ChIP-seq TFs) ES cells. (**C, D**) Bins of miRNAs targeting each gene and their responding Mean CV levels (only interactions with genes in orthologs one2one list have been used) in Mouse and Human ES cells. (**E, F**) Genes (only interactions with genes in orthologs

one2one list have been used) with known protein-protein interactions for Mouse and Human ES cells and their responding Mean CV levels.

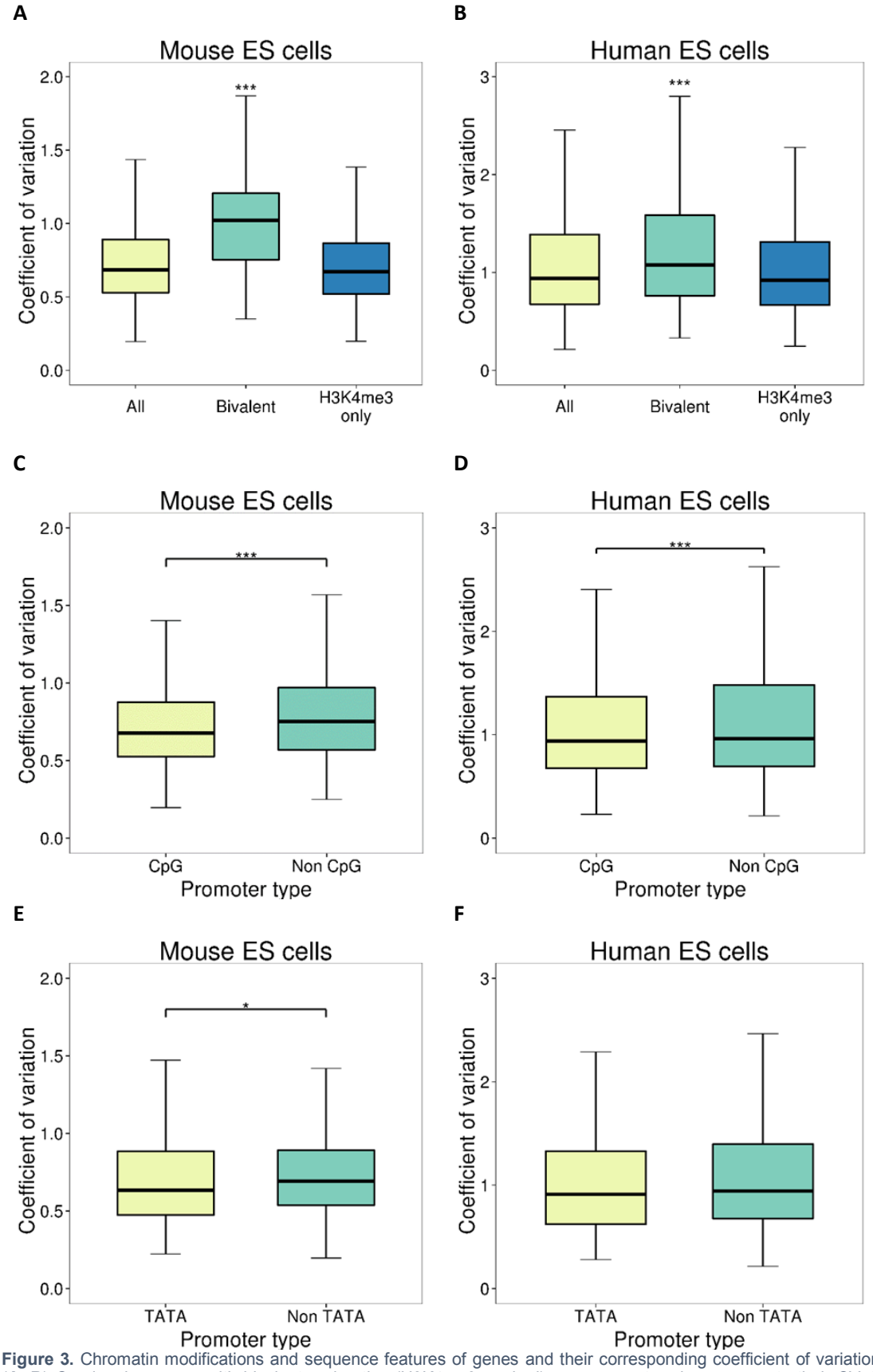


Figure 3. Chromatin modifications and sequence features of genes and their corresponding coefficient of variation. (**A, B**) Overlapping genes with bivalent and active (H3K4me3 marked) gene promoters in response to their CV, in Mouse and Human ES cells. Bivalent genes show significantly higher CV levels than all the promoters (irrespective of overlap) and the active promoters (pairwise t-test, P-value < 0.001) (**C**) CV levels of genes having a CpG island and a non- CpG island promoter. (**D**) CV levels of genes having a TATA box and a non-TATA box promoter.

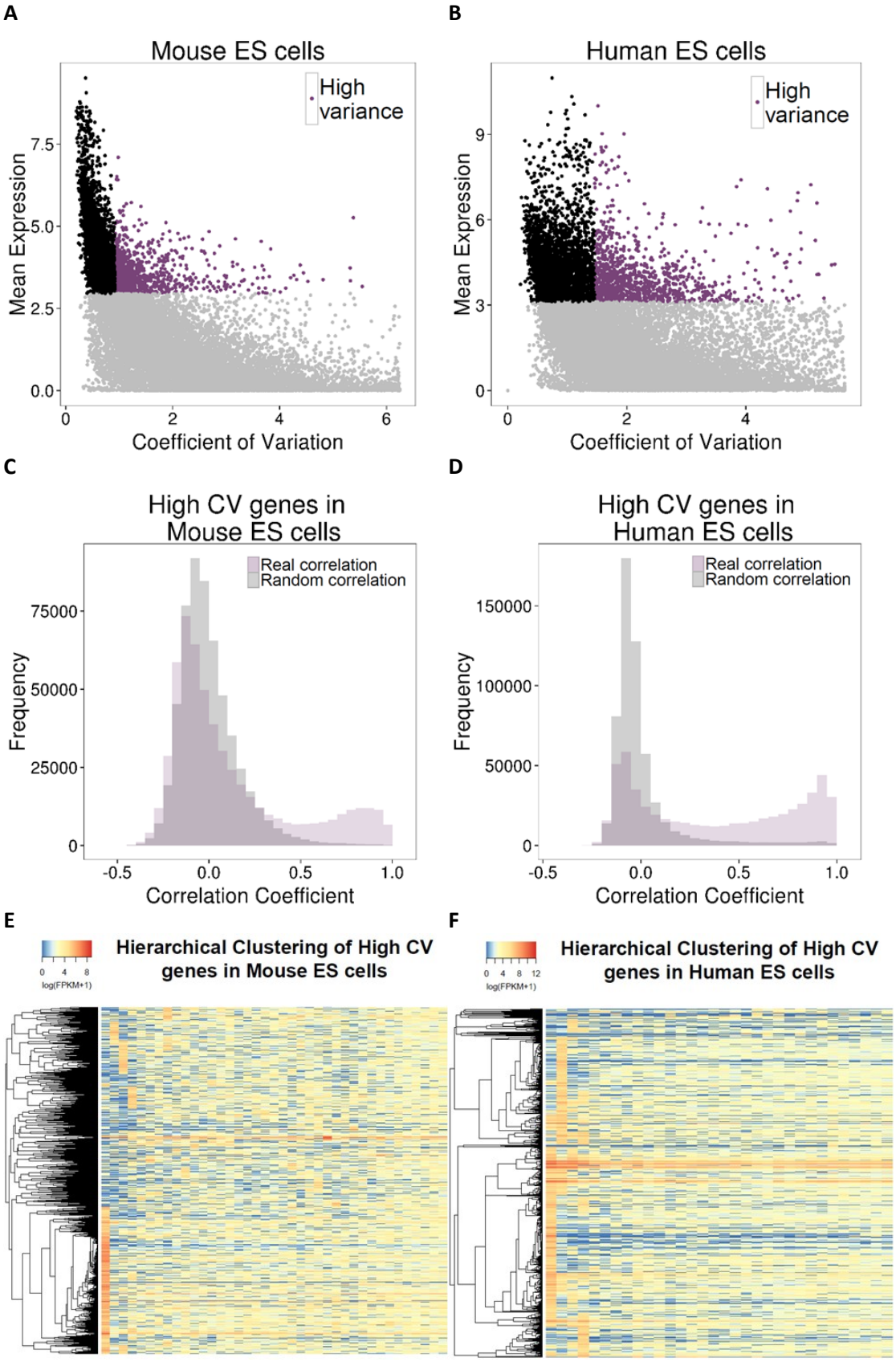


Figure 3. High variance genes are more correlated than expected by chance (**A, B**) Scatterplot of genes in response to their CV and mean expression. Highlighted in purple are the High variance genes, selected based on their CV (CV value greater than the third quartile of the distribution). (**C, D**) Correlation coefficient distributions for the High variance (High CV) genes in Mouse and Human ES cells (statistically significant difference ($p<0.001$, Wilcoxon test) between the real and random distributions). (**E, F**) Heatmaps of gene expression (in log(FPKM+1) values) for the High variance genes (High CV) in Mouse and Human ES cells.

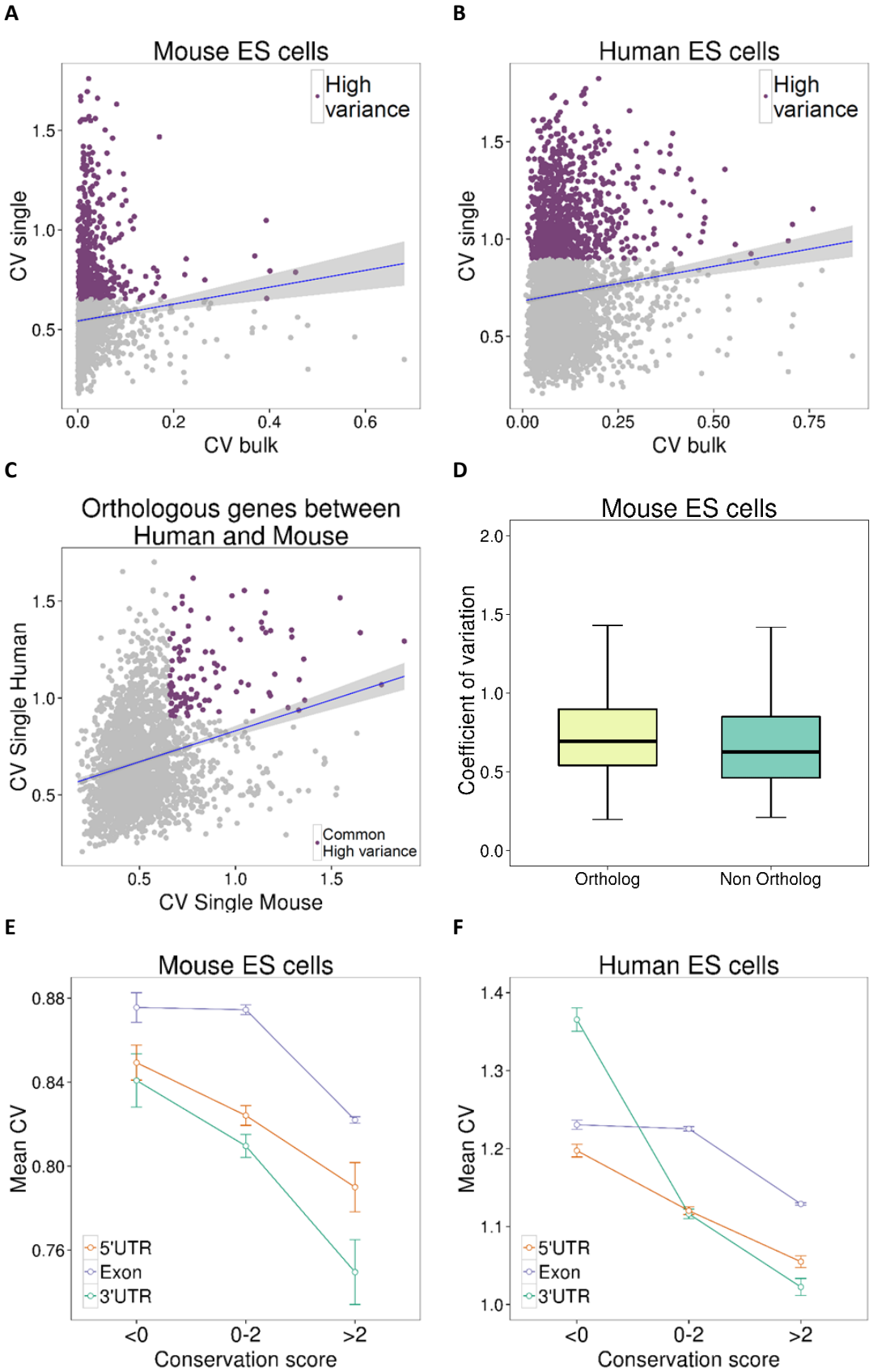


Figure 5. Conservation of expression variability across technologies and species. (**A, B**) Scatterplot of CV values in a bulk expression study against CV values in a single cell RNA–seq study in Mouse and Human ES cells. There is a positive correlation between the CV values of the two technologies (Pearson's *r*=0.06 for mouse, *r*=0.09 for human). (**C**) Scatterplot of CV values of orthologous genes between human and mouse from single RNA-seq studies in ESCs. There is a positive correlation of CV values between species (Pearson's *r*=0.23) and 10% of High CV genes (highlighted in purple) are conserved as highly variant between species (**D**) Boxplot of CV values of orthologous and

non-orthologous genes between human and mouse in ESCs (3,675 orthologs and 554 non-orthologs out of 4,229 genes in our analysis). (**E, F**) Sequence conservation scores and their corresponding Mean CV values for 5'UTR, Exons and 3'UTRs in Mouse and Human ES cells.

Supplementary material for

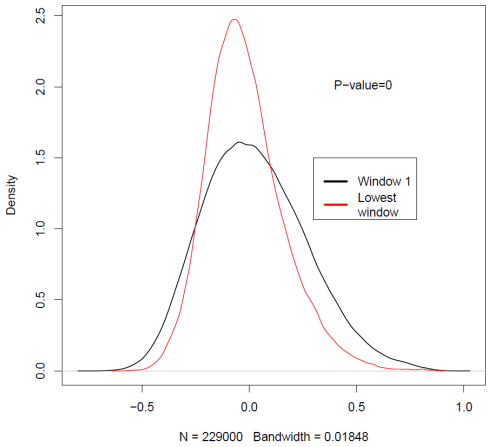
Gene expression variability in mammalian embryonic stem cells using single cell RNA-seq data

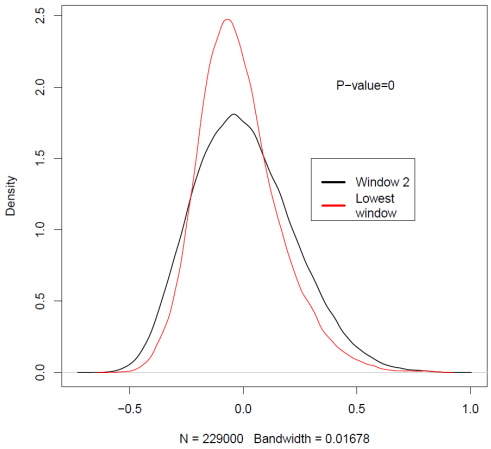
Anna Mantsoki[1], Guillaume Devailly[1], Anagha Joshi[1§]

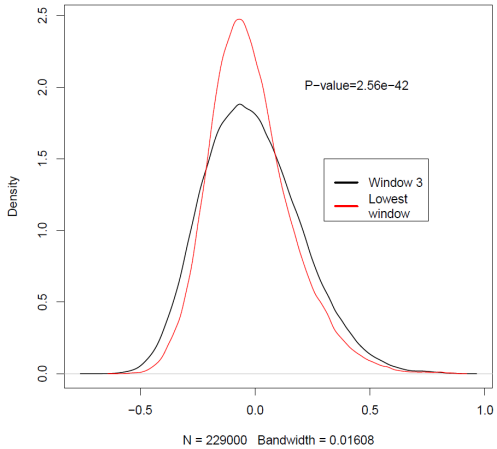
[1]The Roslin institute, University of Edinburgh, Easter bush campus, Midlothian, EH25 9RG.

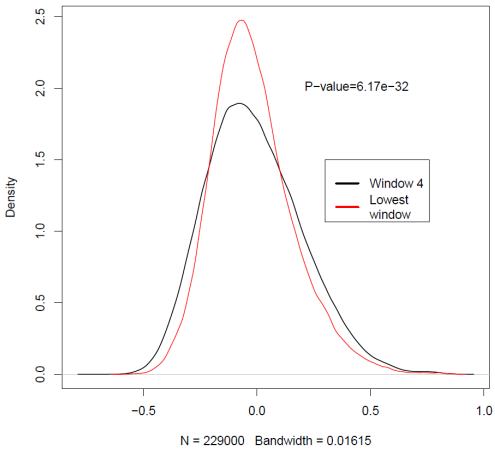
[§]Corresponding author

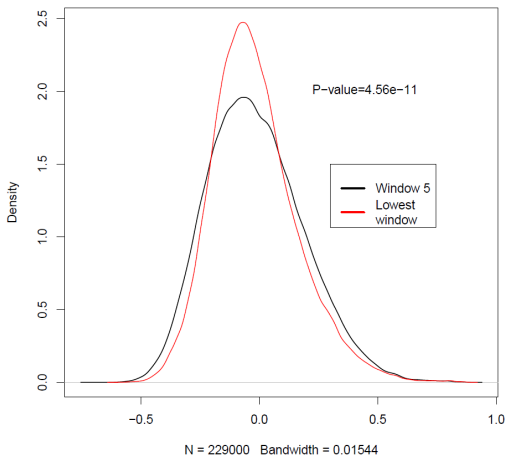
Mouse ES Cells

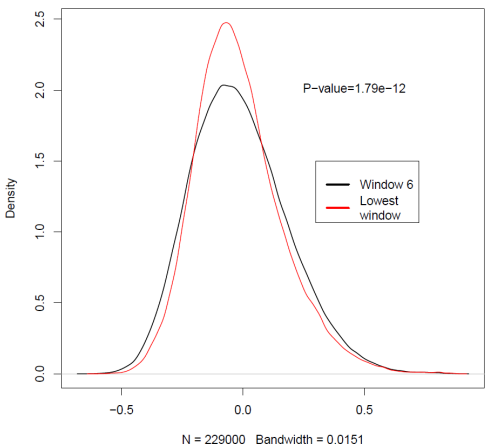


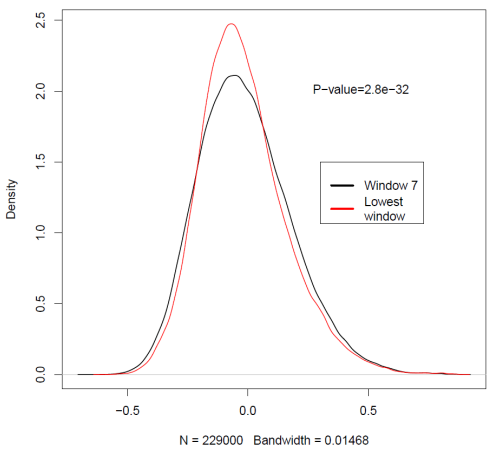


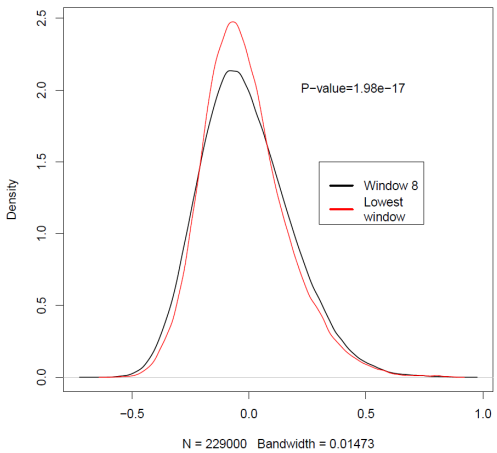


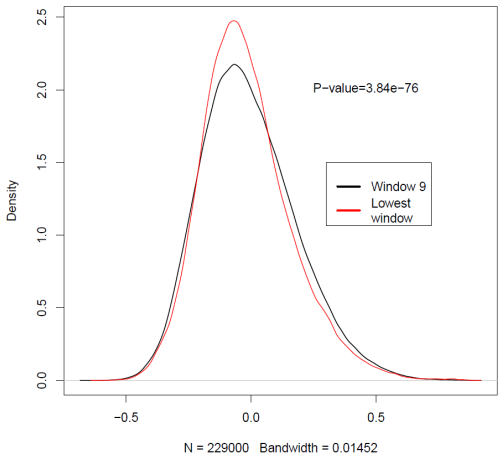


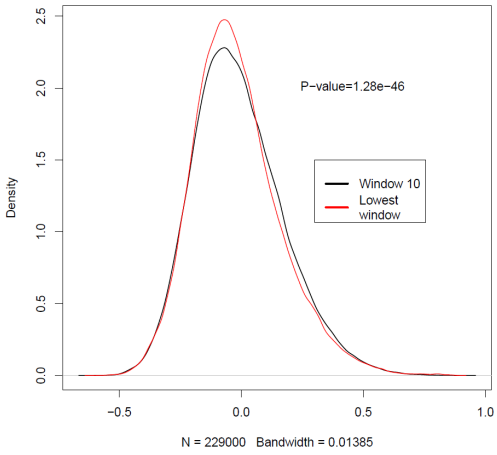


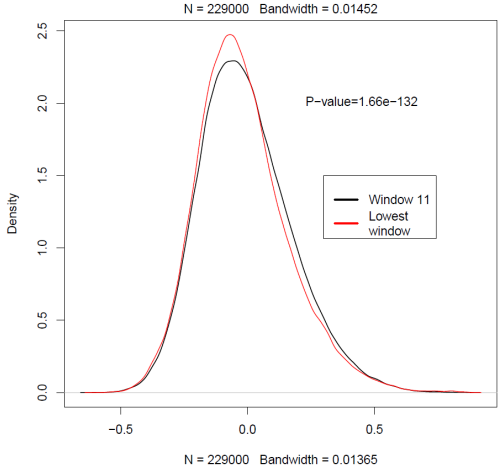


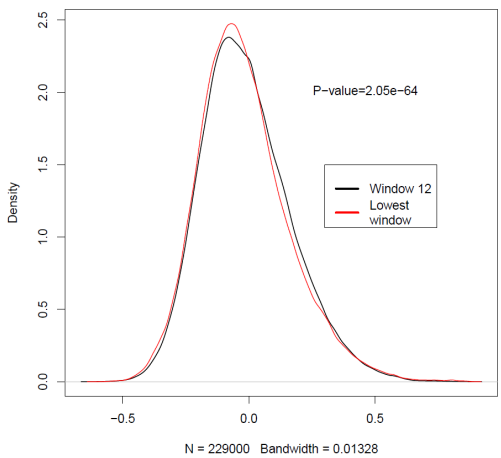


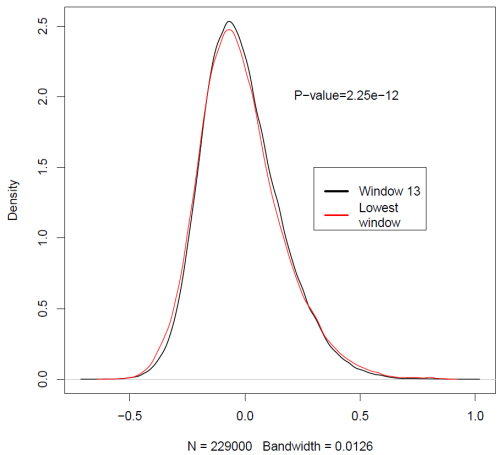


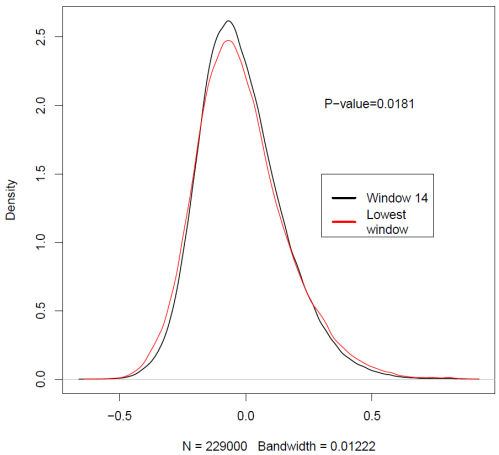

P−value=3.84e−76
Window 9
Lowest window
Density
N = 229000 Bandwidth = 0.01452


P−value=1.28e−46
Window 10
Lowest window
Density
N = 229000 Bandwidth = 0.01385


P−value=1.66e−132
Window 11
Lowest window
Density
N = 229000 Bandwidth = 0.01365


P−value=2.05e−64
Window 12
Lowest window
Density
N = 229000 Bandwidth = 0.01328


P−value=2.25e−12
Window 13
Lowest window
Density
N = 229000 Bandwidth = 0.0126


P−value=0.0181
Window 14
Lowest window
Density
N = 229000 Bandwidth = 0.01222

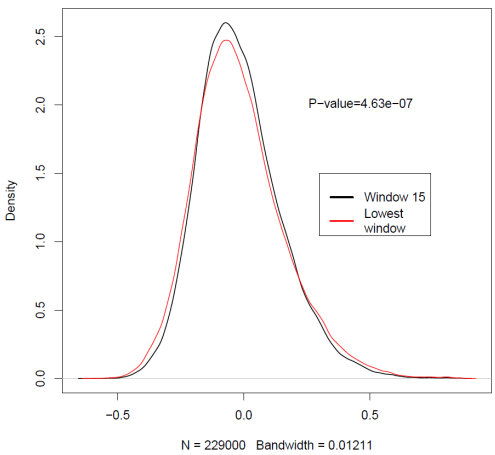
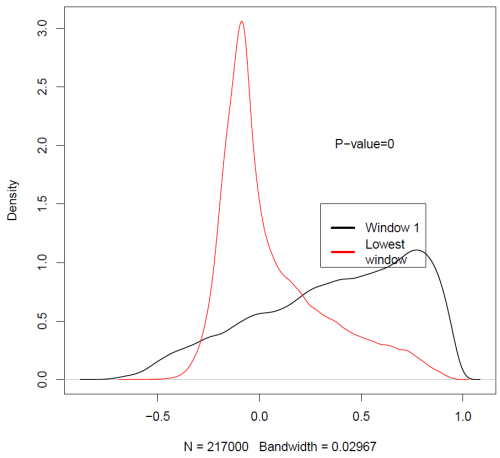
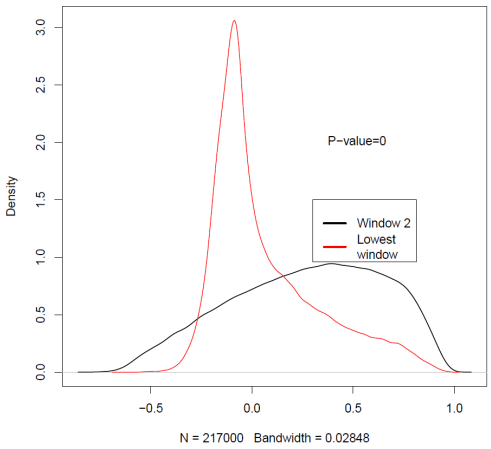
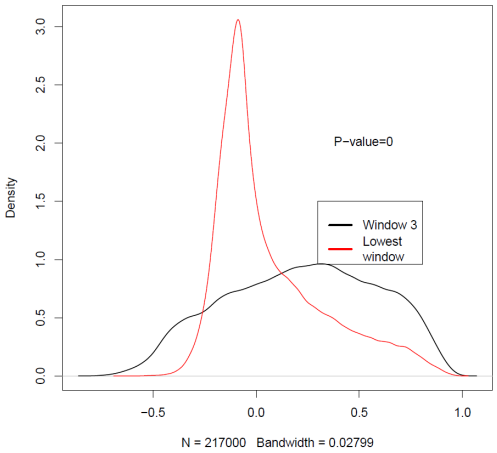

P−value=4.63e−07
Window 15
Lowest window
Density
N = 229000 Bandwidth = 0.01211

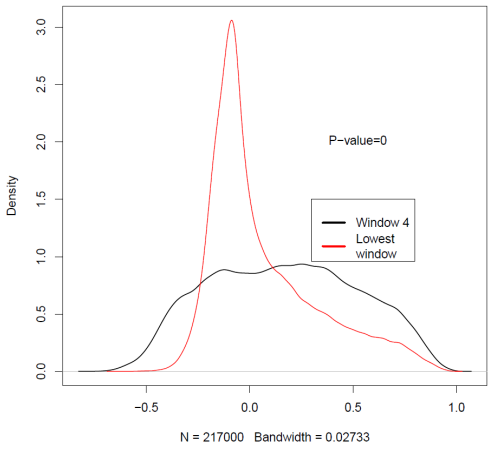
Figure S1. Distributions of correlation of high expression genes with genes in 15 windows of mean expression compared with the lowest mean expression window, in Mouse ES cells. The genes of Windows 1 to 4 were chosen as the ones that are above the threshold of technical variation.

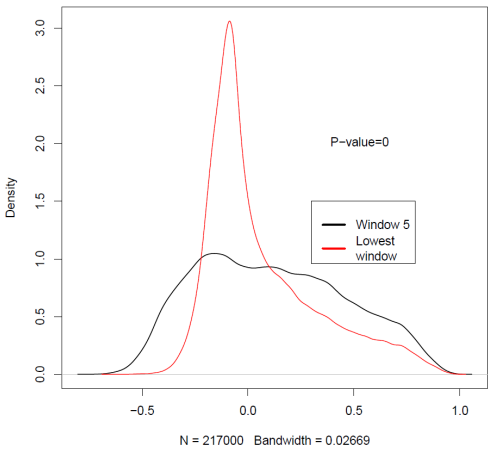
Human ES Cells

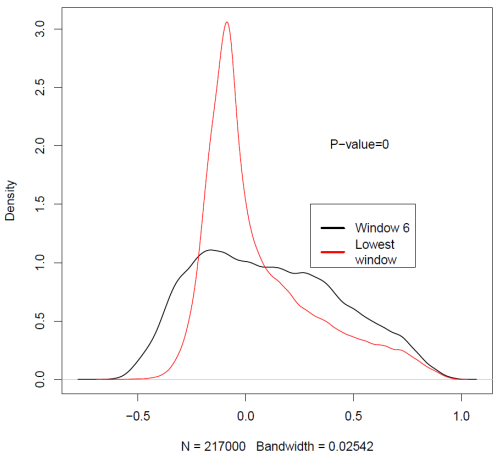


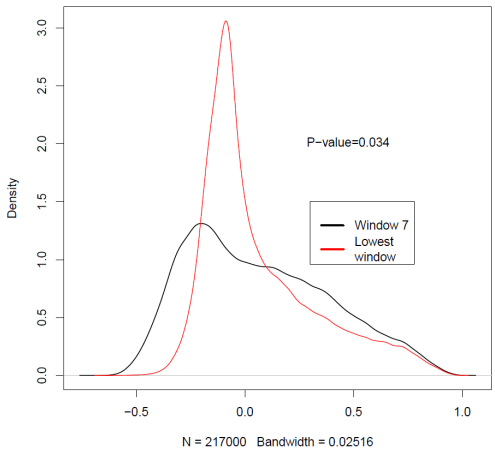


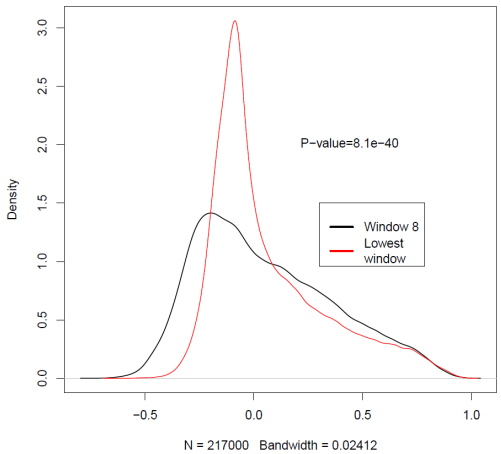


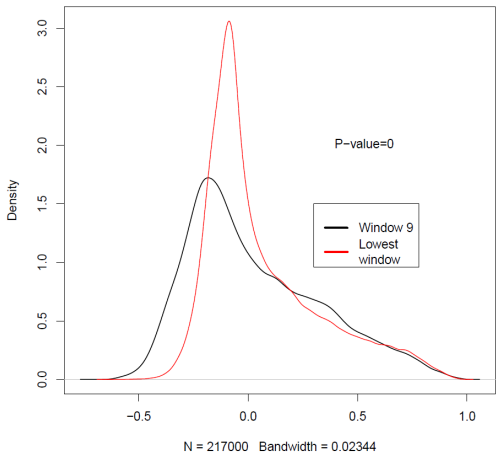


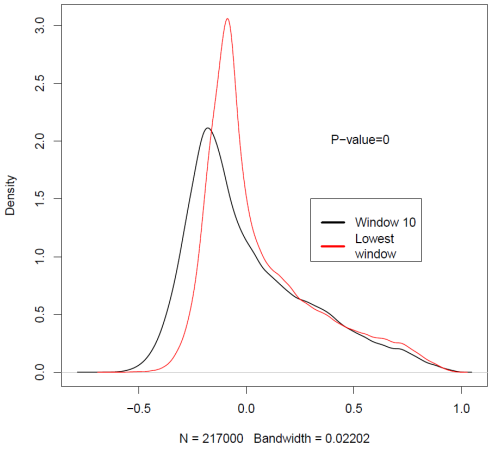


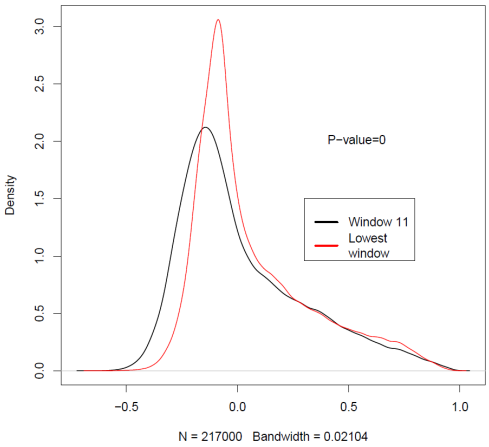


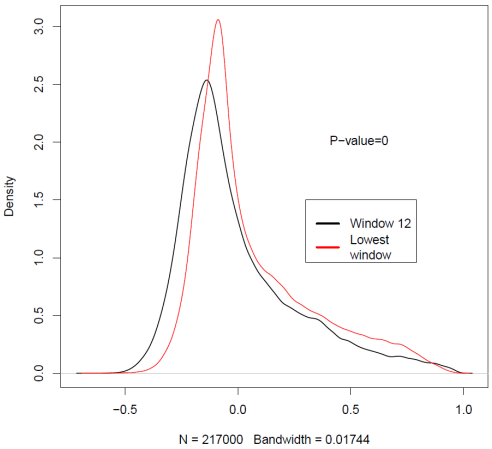


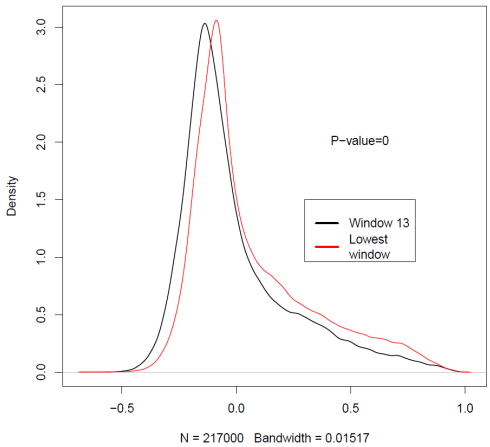


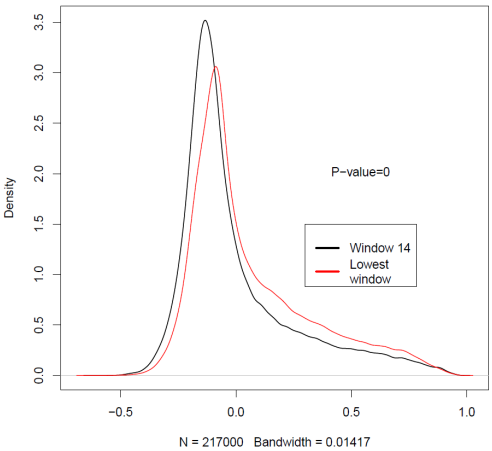

Density
P−value=0
Window 9
Lowest
window
N = 217000 Bandwidth = 0.02344


Density
P−value=0
Window 10
Lowest
window
N = 217000 Bandwidth = 0.02202


Density
P−value=0
Window 11
Lowest
window
N = 217000 Bandwidth = 0.02104


Density
P−value=0
Window 12
Lowest
window
N = 217000 Bandwidth = 0.01744


Density
P−value=0
Window 13
Lowest
window
N = 217000 Bandwidth = 0.01517


Density
P−value=0
Window 14
Lowest
window
N = 217000 Bandwidth = 0.01417

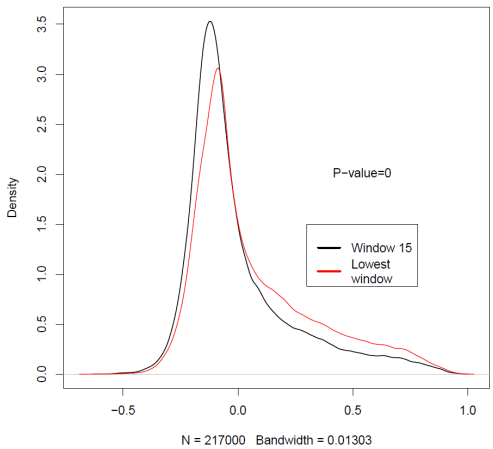

Density
P−value=0
Window 15
Lowest
window
N = 217000 Bandwidth = 0.01303

Figure S2. Distributions of correlation of high expression genes with genes in 18 windows of mean expression compared with the lowest mean expression window, in Human ES cells. The genes of Windows 1 to 4 were chosen as the ones that are above the threshold of technical variation.

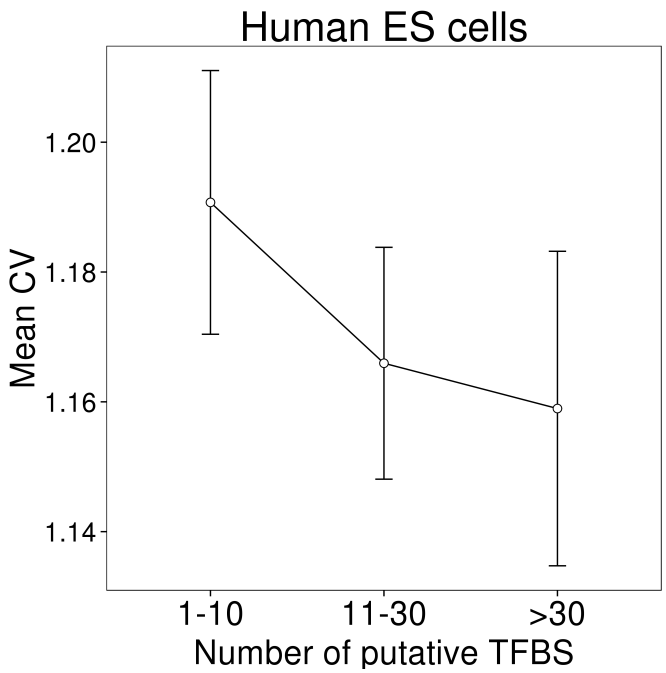


Figure S3. Number of putative Transcription Factor Binding Sites (TFBS) per gene (shown in 3 bins) and their corresponding Mean CV values. There was no statistically significant difference between means

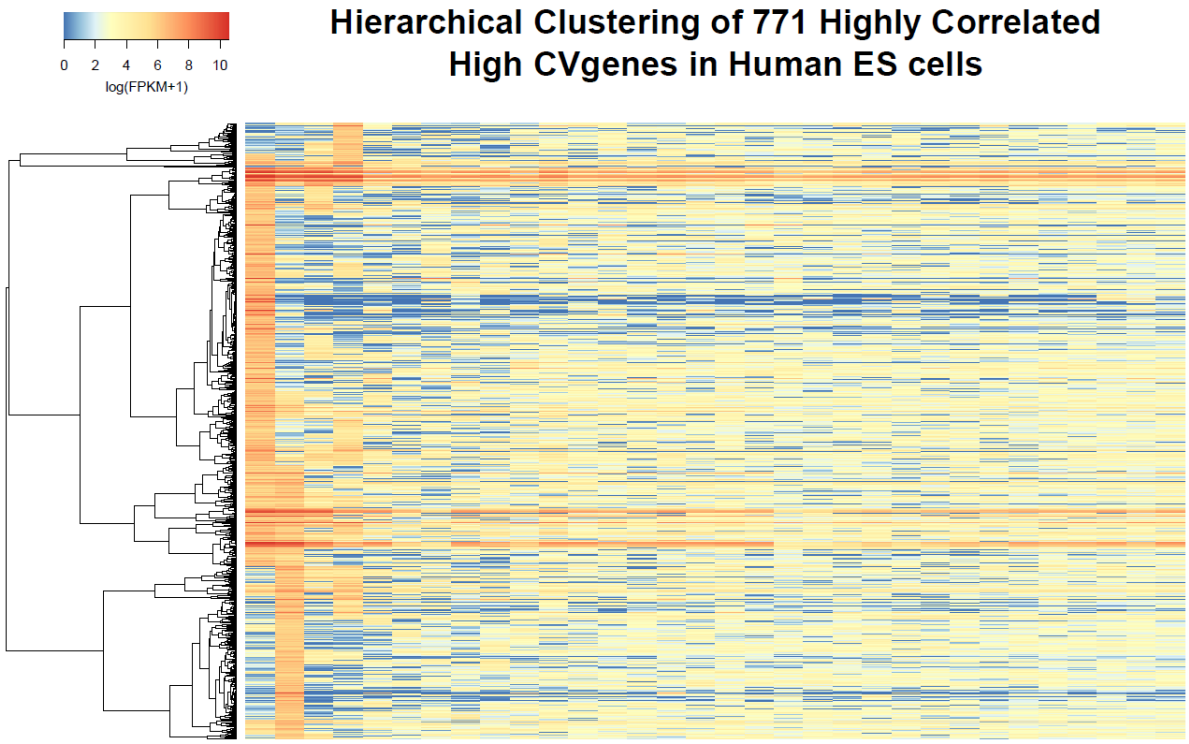


Figure S4. Heatmap of gene expression values (in log(FPKM+1)) of highly correlated variable (High CV) genes in Mouse ES cells.

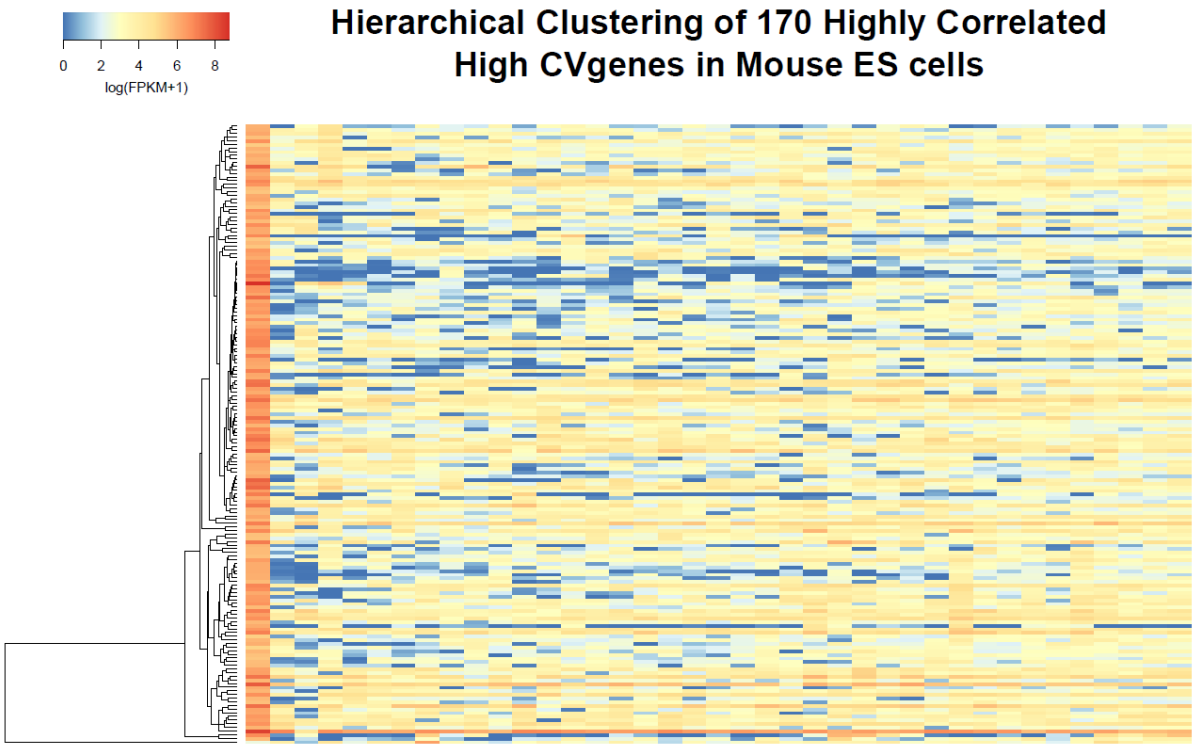


Figure S5. Heatmap of gene expression values (in log(FPKM+1)) of highly correlated variable (High CV) genes in Human ES cells.

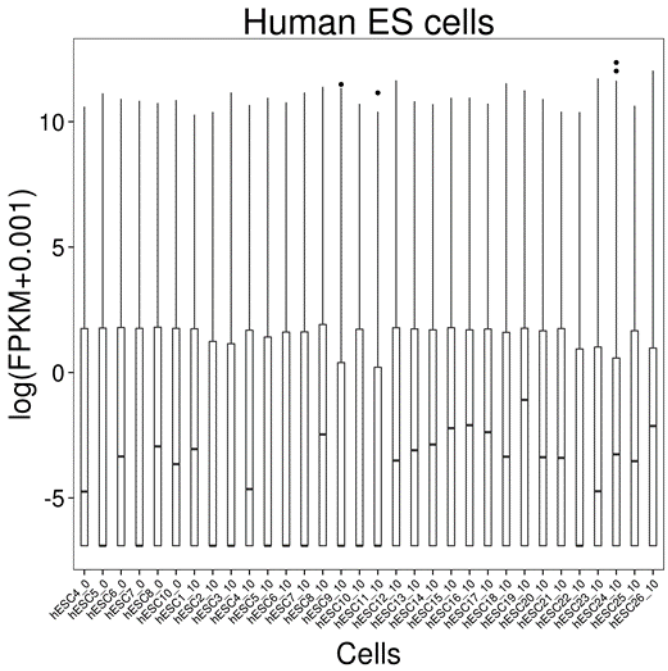


Figure S6. Boxplot of FPKM values for all cells in human

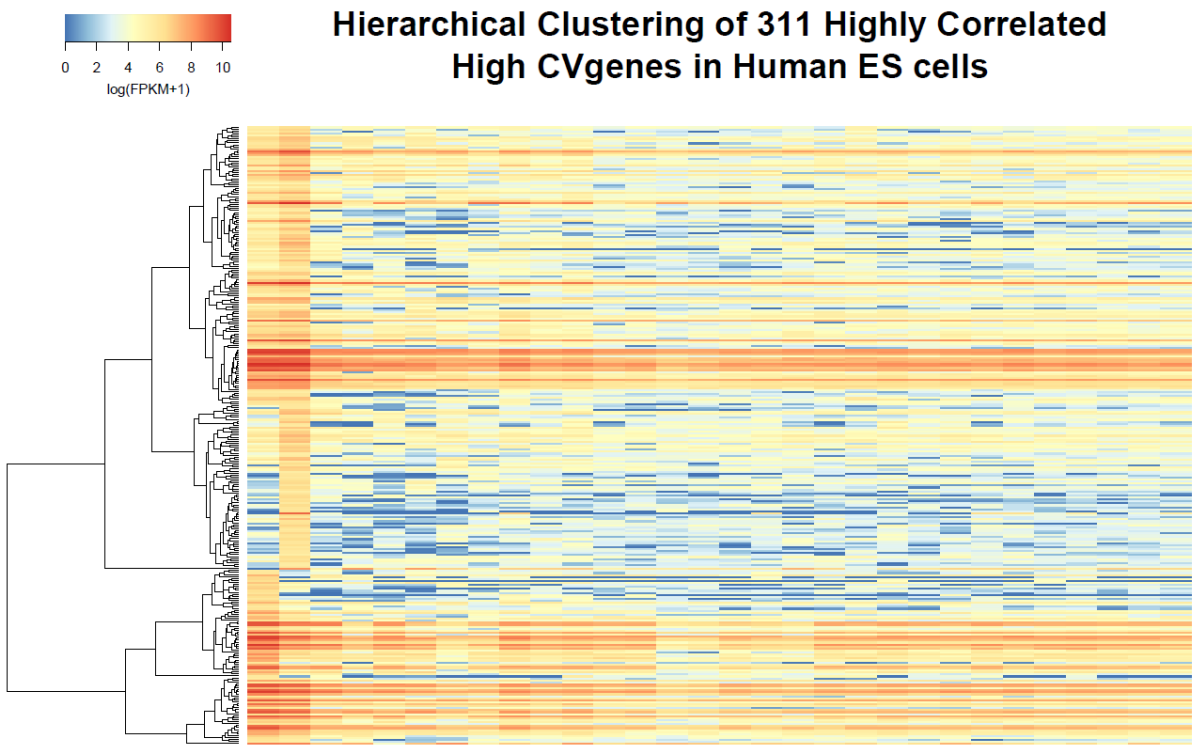


Figure S7. Heatmap of High CV genes in Human ES cells after discarding cells 24 and 26